Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy

Hanchuan Peng [*], and Fuhui Long [*]
Lawrence Berkeley National Lab, Berkeley, CA, 94720. Email: {hpeng, flong}@lbl.gov

[*] Both authors contribute equally to this work.

Abstract *- Feature selection is an important problem for pattern classification systems. We study how to select good features according to the maximal statistical dependency criterion based on mutual information. Because of the difficulty in directly implementing the maximal dependency condition, we first derive an equivalent form, called minimal-redundancy-maximal-relevance criterion (mRMR), for first-order incremental feature selection. Then, we present a two-stage feature selection algorithm by combining mRMR and other more sophisticated feature selectors (e.g. wrappers). This allows us to select a compact set of superior features at very low cost. We perform extensive experimental comparison of our algorithm and other methods using three different classifiers (naïve Bayes, support vector machine, and linear discriminate analysis) and four different data sets (handwritten digits, arrhythmia, NCI cancer cell lines, and lymphoma tissues). The results confirm that mRMR leads to promising improvement on feature selection and classification accuracy.*

Index Terms - Feature selection, mutual information, minimal redundancy, maximal relevance, maximal dependency, classification

1. Introduction

In many pattern recognition applications, identifying the most characterizing features (or attributes) of the observed data, i.e., feature selection (or variable selection, among many other names) [30][14][17][18][15][12][11][19][31][32][5], is critical to minimize the classification error. Given the input data D tabled as N samples and M features $X = \{x_i, i=1,\ldots,M\}$, and the target classification variable c, feature selection problem is to find from the M-dimensional observation space, R^M, a subspace of m features, R^m, that "optimally" characterizes c.

Given a condition defining the "optimal characterization", a search algorithm is needed to find the best subspace. Because the total number of subspaces is 2^M, and the number of subspaces with dimensions no larger than m is $\Sigma_{i=1}^{m}(\begin{smallmatrix}M\\ i\end{smallmatrix})$, it is hard to search the feature subspace exhaustively. Alternatively, many sequential-search based approximation schemes have been proposed, including best individual features, sequential forward search, sequential forward floating search, etc (see [30][14][13] for detailed comparison.).

The optimal characterization condition often means the *minimal classification error*. In an unsupervised situation where the classifiers are not specified, minimal error usually requires the maximal statistical dependency of the target class c on the data distribution in the subspace R^m (and vice versa). This scheme is *maximal dependency* (Max-Dependency).

One of the most popular approaches to realize Max-Dependency is *maximal relevance* (Max-Relevance) feature selection: selecting the features with the highest relevance to the target class c. Relevance is usually characterized in terms of correlation or mutual information, of which the latter is one of the widely used measures to define dependency of variables. In this paper, we focus on the discussion of mutual information based feature selection.

Given two random variables x and y, their mutual information is defined in terms of their probabilistic density functions $p(x)$, $p(y)$ and $p(x,y)$:

$$I(x;y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy\,. \qquad (1)$$

In Max-Relevance, the selected features x_i are required, individually, to have the largest mutual information $I(x_i; c)$ with the target class c, reflecting the largest dependency on the target class. In terms of sequential search, the m best individual features, i.e. the top m features in the descent ordering of $I(x_i; c)$, are often selected as the m features.

In feature selection, it has been recognized that the combinations of individually good features do not necessarily lead to good classification performance. In other words, "the m best features are not the best m features" [4][3][14][30]. Some researchers have studied indirect or direct means to reduce the redundancy among features[1] (e.g. [4][14][19][15][22][12][5]) and select features with the *minimal redundancy* (Min-Redundancy). For example, in the sequential forward floating search [25], the joint dependency of features on the target class is maximized; as a by-product, the redundancy among features might be reduced. In [12], Jaeger *et al* presented a pre-filtering method to group variables, thus redundant variables within each group can be removed. In [5], we proposed a heuristic *minimal-redundancy-maximal-relevance* (mRMR) framework to minimize redundancy, and used a series of intuitive measures of relevance and redundancy to select promising features for both continuous and discrete data sets.

Our work in this paper focuses on three issues that have not been touched in earlier work. First, although both Max-Relevance and Min-Redundancy have been intuitively used for feature selection, no theoretical analysis is given on why they can benefit selecting optimal features for classification. Thus the first goal of this paper is to present a theoretical analysis showing that mRMR is equivalent to Max-Dependency for first-order feature selection, but is more efficient.

Second, we investigate how to combine mRMR with other feature selection methods (such as wrappers [18][15]) into a two-stage selection algorithm. By doing this, we show that the space of candidate features selected by mRMR is more characterizing. This property of mRMR facilitates the integration of other feature selection schemes to find a compact subset of superior features at very low cost.

Third, through comprehensive experiments we compare mRMR, Max-Relevance, Max-Dependency, and the two-stage feature selection algorithm, using three different classifiers and four data sets. The results show that mRMR and our two-stage algorithm are very effective in a wide range of feature selection applications.

This paper is organized as follows. Section 2 presents the theoretical analysis of the relationships of Max-Dependency, Max-Relevance, and Min-Redundancy. Section 3 presents the two-stage feature selection algorithm, including schemes to integrate wrappers to select a squeezed subset of features. Section 4 discusses implementation issues of density estimation for mutual information, and several different classifiers. Section 5 gives experimental results on four data sets, including handwritten characters, arrhythmia, NCI cancer cell lines, and lymphoma tissues. Sections 6 and 7 are discussions and conclusions, respectively.

2. Relationships of Max-Dependency, Max-Relevance and Min-Redundancy

2.1 Max-Dependency

In term of mutual information, the purpose of feature selection is to find a feature set S with m features $\{x_i\}$, which jointly have the largest dependency on the target class c. This scheme, called Max-Dependency, has the following form:

$$\max D(S,c), \quad D = I(\{x_i, i = 1,...,m\}; c). \tag{2}$$

Obviously, when m equals 1, the solution is the feature that maximizes $I(x_j; c)$ ($1 \le j \le M$). When $m > 1$, a simple incremental search scheme is to add one feature at one time: given the set with $m-1$ features, S_{m-1}, the mth feature can be determined as the one that contributes to the largest increase of $I(S; c)$, which takes the form of Eq. (3).

[1] Minimal redundancy has also been studied in feature extraction, which aims to find good features in a transformed domain. For instance, it has been well addressed in various techniques such as principal component analysis and independent component analysis [10], neural network feature extractors (e.g. [22]), etc.

$$I(S_m;c) = \iint p(S_m,c)\log\frac{p(S_m,c)}{p(S_m)p(c)}dS_m dc$$
$$= \iint p(S_{m-1},x_m,c)\log\frac{p(S_{m-1},x_m,c)}{p(S_{m-1},x_m)p(c)}dS_{m-1}dx_m dc \quad (3)$$
$$= \int\cdots\int p(x_1,\cdots,x_m,c)\log\frac{p(x_1,\cdots,x_m,c)}{p(x_1,\cdots,x_m)p(c)}dx_1\cdots dx_m dc.$$

Despite the theoretical value of Max-Dependency, it is often hard to get an accurate estimation for multivariate density $p(x_1, \ldots, x_m)$ and $p(x_1, \ldots, x_m, c)$, because of two difficulties in the high-dimensional space: 1) the number of samples is often insufficient, and 2) the multivariate density estimation often involves computing the inverse of the high-dimension covariance matrix, which is usually an ill-posed problem. Another drawback of Max-Dependency is the slow computational speed. These problems are most pronounced for continuous feature variables.

Even for discrete (categorical) features, the practical problems in implementing Max-Dependency cannot be completely avoided. For example, suppose each feature has 3 categorical states and N samples. K features could have a maximum $\min(3^K, N)$ joint states. When the number of joint states increases very quickly and gets comparable to the number of samples, N, the joint probability of these features, as well as the mutual information, cannot be estimated correctly. Hence, although Max-Dependency feature selection might be useful to select a very small number of features when N is large, it is not appropriate for applications where the aim is to achieve high classification accuracy with a reasonably compact set of features.

2.2 Max-Relevance and Min-Redundancy

As Max-Dependency criterion is hard to implement, an alternative is to select features based on *maximal relevance* criterion (Max-Relevance). Max-Relevance is to search features satisfying Eq. (4), which approximates $D(S,c)$ in Eq. (2) with the mean value of all mutual information values between individual feature x_i and class c.

$$\max D(S,c), \quad D = \frac{1}{|S|}\sum_{x_i\in S} I(x_i;c). \quad (4)$$

It is likely that features selected according to Max-Relevance could have rich redundancy, i.e. the dependency among these features could be large. When two features highly depend on each other, the respective class-discriminative power would not change much if one of them were removed. Therefore, the following *minimal redundancy* (Min-Redundancy) condition can be added to select mutually exclusive features [5]:

$$\min R(S), \quad R = \frac{1}{|S|^2}\sum_{x_i,x_j\in S} I(x_i,x_j). \quad (5)$$

The criterion combining the above two constraints is called "*minimal-redundancy-maximal-relevance*" (mRMR) [5]. We define the operator $\Phi(D, R)$ to combine D and R and consider the following simplest form to optimize D and R simultaneously:

$$\max \Phi(D,R), \quad \Phi = D - R. \quad (6)$$

In practice, incremental search methods can be used to find the near-optimal features defined by $\Phi(.)$. Suppose we already have S_{m-1}, the feature set with $m-1$ features. The task is to select the mth feature from the set $\{X - S_{m-1}\}$. This is done by selecting the feature that maximizes $\Phi(.)$. The respective incremental algorithm optimizes the following condition:

$$\max_{x_j\in X-S_{m-1}}\left[I(x_j;c) - \frac{1}{m-1}\sum_{x_i\in S_{m-1}} I(x_j;x_i)\right]. \quad (7)$$

The computational complexity of this incremental search method is $O(|S|\cdot M)$.

2.3 Optimal First-Order Incremental Selection

We prove in the following that the combination of Max-Relevance and Min-Redundancy criteria, i.e., the mRMR criterion, is equivalent to the Max-Dependency criterion if one feature is selected (added) at one time. We call this type of selection the "first-order" incremental search. We have the following theorem.

Theorem: For the first-order incremental search, mRMR is equivalent to Max-Dependency (Eq.(2)).

Proof: By definition of the first-order search, we assume that S_{m-1}, i.e. the set of m-1 features, has already been obtained. The task is to select the optimal mth feature x_m from set $\{X - S_{m-1}\}$.

The dependency D in Eqs.(2) and (3) is represented by mutual information, i.e., $D = I(S_m; c)$ where $S_m = \{S_{m-1}, x_m\}$ can be treated as a multivariate variable. Thus by the definition of mutual information, we have:

$$\begin{aligned} I(S_m;c) &= H(c) + H(S_m) - H(S_m, c) \\ &= H(c) + H(S_{m-1}, x_m) - H(S_{m-1}, x_m, c). \end{aligned} \tag{8}$$

where $H(.)$ is the entropy of the respective multivariate (or univariate) variables.

Now we define the following quantity $J(S_m) = J(x_1, \ldots, x_m)$ for scalar variables $x_1, \ldots, x_m$,

$$J(x_1, x_2, ..., x_m) = \int \cdots \int p(x_1, ..., x_m) \log \frac{p(x_1, x_2, ..., x_m)}{p(x_1) \cdots p(x_m)} dx_1 \cdots dx_m \,. \tag{9}$$

Similarly, we define $J(S_m, c) = J(x_1, \ldots, x_m, c)$ as

$$J(x_1, x_2, ..., x_m, c) = \int \cdots \int p(x_1, ..., x_m, c) \log \frac{p(x_1, x_2, ..., x_m, c)}{p(x_1) \cdots p(x_m) p(c)} dx_1 \cdots dx_m dc \,. \tag{10}$$

We can easily derive Eqs. (11) and (12) from Eqs.(9) and (10),

$$H(S_{m-1}, x_m) = H(S_m) = \Sigma_{i=1}^{m} H(x_i) - J(S_m), \tag{11}$$

$$H(S_{m-1}, x_m, c) = H(S_m, c) = H(c) + \Sigma_{i=1}^{m} H(x_i) - J(S_m, c). \tag{12}$$

By substituting them to the corresponding terms in Eq. (8), we have

$$\begin{aligned} I(S_m, c) &= J(S_m, c) - J(S_m) \\ &= J(S_{m-1}, x_m, c) - J(S_{m-1}, x_m). \end{aligned} \tag{13}$$

Obviously, Max-Dependency is equivalent to simultaneously maximizing the first term and minimizing the second term.

We can use the Jensen's Inequality [16] to show the second term $J(S_{m-1}, x_m)$ is lower-bounded by 0. A related and slightly simpler proof is to consider the inequality $\log(z) \le z - 1$ with the equality if and only if $z = 1$. We see that

$$\begin{aligned} -J(x_1, x_2, ..., x_m) &= \int \cdots \int p(x_1, ..., x_m) \log \frac{p(x_1) \cdots p(x_m)}{p(x_1, ..., x_m)} dx_1 \cdots dx_m \\ &\le \int \cdots \int p(x_1, ..., x_m) [\frac{p(x_1) \cdots p(x_m)}{p(x_1, ..., x_m)} - 1] dx_1 \cdots dx_m \\ &= \int \cdots \int p(x_1) \cdots p(x_m) dx_1 \cdots dx_m - \int \cdots \int p(x_1, \cdots, x_m) dx_1 \cdots dx_m \\ &= 1 - 1 = 0. \end{aligned} \tag{14}$$

It is easy to verify that the minimum is attained when $p(x_1, \ldots, x_m) = p(x_1) \cdots p(x_m)$, i.e. all the variables are independent of each other. As all the m–1 features have been selected, this pair-wise independence condition means that the mutual information between x_m and any selected feature x_i (i = 1,…,m–1) is minimized. This is the Min-Redundancy criterion.

We can also derive the upper bound of the first term in Eq. (13), $J(S_{m-1}, c, x_m)$. For simplicity, let's first show the upper bound of the general form $J(y_1, ..., y_n)$, assuming there are n variables $y_1, \ldots, y_n$. This can be seen as follows.

$$\begin{aligned} J(y_1, y_2, ..., y_n) &= \int \cdots \int p(y_1, ..., y_n) \log \frac{p(y_1, ..., y_n)}{p(y_1) \cdots p(y_n)} dy_1 \cdots dy_n \\ &= \int \cdots \int p(y_1, ..., y_n) \log \frac{p(y_1 \mid y_2, ..., y_n) p(y_2 \mid y_3, ..., y_n) \cdots p(y_{n-1} \mid y_n) p(y_n)}{p(y_1) \cdots p(y_{n-1}) p(y_n)} dy_1 \cdots dy_n \\ &= \sum_{i=1}^{n-1} H(y_i) - H(y_1 \mid y_2, ..., y_n) - H(y_2 \mid y_3, ..., y_n) - H(y_{n-1} \mid y_n) \\ &\le \sum_{i=1}^{n-1} H(y_i). \end{aligned} \tag{15}$$

Eq. (15) can be easily extended as

$$J(y_1, y_2, ..., y_n) \le \min\{\sum_{i=2}^{n} H(y_i), \sum_{i=1, i\ne 2}^{n} H(y_i), \cdots, \sum_{i=1, i\ne n-1}^{n} H(y_i), \sum_{i=1}^{n-1} H(y_i)\}. \tag{16}$$

It is easy to verify the maximum of $J(y_1, ..., y_n)$, or similarly the first term in Eq. (13), $J(S_{m-1}, c, x_m)$, is attained when all variables are maximally dependent. When S_{m-1} has been fixed, this indicates that x_m and c should have the maximal dependency. This is the Max-Relevance criterion.

Therefore, according to Eq. (13), as a combination of Max-Relevance and Min-Redundancy, mRMR is equivalent to Max-Dependency for first-order selection.

□

Note that the quantity $J(.)$ in Eqs. (9) and (10) has also been called "mutual information" for multiple scalar variables [10]. We have the following observations:

1) Minimizing $J(S_m)$ only is equivalent to searching mutually exclusive (independent) features[2]. This is insufficient for selecting highly discriminative features.
2) Maximizing $J(S_m, c)$ only leads to Max-Relevance. Clearly, the difference between mRMR and Max-Relevance is rooted in the different definitions of dependency (in term of mutual information). Eq.(10) does not consider the joint effect of features on the target class. On the contrary, Max-Dependency (Eqs. (2) and (3)) considers the dependency between the data distribution in subspace R^m and the target class c. This difference is critical in many circumstances.
3) The equivalence between Max-Dependency and mRMR indicates mRMR is an *optimal* first-order implementation scheme of Max-Dependency.
4) Compared to Max-Dependency, mRMR avoids the estimation of multivariate densities $p(x_1, \ldots, x_m)$ and $p(x_1, \ldots, x_m, c)$. Instead, calculating the bivariate density $p(x_i, x_j)$ and $p(x_i, c)$ could be much easier and more accurate. This also leads to a more efficient feature selection algorithm.

3. Feature Selection Algorithms

Our goal is to design efficient algorithms to select a compact set of features. In Section 2 we propose a fast mRMR feature selection scheme (Eq. (7)). A remaining issue is how to determine the optimal number of features m. Since a mechanism to remove potentially redundant features from the already selected features has not been considered in the incremental selection, according to the idea of mRMR we need to refine the results of incremental selection.

We present a two-stage feature selection algorithm. In the first stage, we find a candidate feature set using the mRMR incremental selection method. In the second stage, we use other more sophisticated schemes to search a compact feature subset from the candidate feature set.

[2] In the field of feature extraction, minimizing $J(S_m)$ has led to an algorithm of independent component analysis [10].

3.1 Selecting the Candidate Feature Set

To select the candidate feature set, we compute the cross-validation classification error for a large number of features and find a relatively stable range of small error. This range is called Ω. The optimal number of features (denoted as n^*) of the candidate set is determined within Ω. The whole process includes three steps:

1) Use mRMR incremental selection (Eq.(7)) to select n (a preset large number) sequential features from the input X. This leads to n sequential feature sets $S_1 \subset S_2 \subset \ldots \subset S_{n-1} \subset S_n$.
2) Compare all the n sequential feature sets $S_1, \ldots, S_k, \ldots, S_n$, $(1 \le k \le n)$ to find the range of k, called Ω, within which the respective (cross-validation classification) error e_k is consistently small (i.e. has both small mean and small variance).
3) Within Ω, find the smallest classification error $e^* = \min e_k$. The optimal size of the candidate feature set, n^*, is chosen as the smallest k that corresponds to e^*.

3.2 Selecting Compact Feature Subsets

Many sophisticated schemes can be used to search the compact feature subsets from the candidate set S_{n^*}. To illustrate that mRMR can produce better candidate features, which favors better combination with other methods, we use wrappers to search the compact feature subsets.

A wrapper [15][18] is a feature selector that convolves with a classifier (e.g. naïve Bayes classifier), with the direct goal to minimize the classification error of the particular classifier. Usually, wrappers could yield high classification accuracy for a particular classifier, at the cost of high computational complexity and less generalization of the selected features on other classifiers. This is different from the mRMR method introduced above, which does not optimize the classification error directly. The latter type of approach (e.g. mRMR and Max-Relevance), sometimes called "filter" [18][15], often selects features by testing whether some preset conditions about the features and the target class are satisfied. In practice, the filter approach has much lower complexity than wrappers; the features thus selected often yield comparable classification errors for different classifiers, because such features often form intrinsic clusters in the respective subspace.

By using mRMR feature selection in the first-stage, we intend to find a small set of candidate features, on which the wrappers can be applied at a much lower cost in the second-stage. We will continue our discussion on this point in Section 3.3.

In this paper, we consider two selection schemes of wrapper, i.e. the backward and forward selections:

1) The backward selection tries to exclude one redundant feature at a time from the current feature set S_k (initially, k is set to n^* obtained in Section 3.1), with the constraint that the resultant feature set S_{k-1} leads to a classification error e_{k-1} no worse than e_k. Because every feature in S_k can be considered in removal, there are k different configurations of S_{k-1}. For each possible configuration, the respective classification error e_{k-1} is calculated. If for every configuration the corresponding e_{k-1} is larger than e_k, there is no gain in either classification accuracy or feature dimension reduction (i.e. every existing feature in S_k appears to be useful), thus the backward selection terminates (accordingly, the size of the compact feature subset, m, is set to k). Otherwise, among the k configurations of S_{k-1}, the one that leads to the largest error reduction is chosen as the new feature set. If there are multiple configurations leading to the same error reduction, one of them is chosen randomly. This decremental selection procedure is repeated until the termination condition is satisfied.
2) The forward selection tries to select a subset of m features from S_{n^*} in an incremental manner. Initially the classification error is set to the number of samples, i.e. N. The wrapper first searches for the feature subset with one feature, denoted as Z_1, by selecting the feature x_1^* that leads to the largest error reduction. Then from the set $\{S_n - Z_1\}$ the wrapper selects the feature

x_2^* so that the feature set $Z_2 = \{Z_1, x_2^*\}$ leads to the largest error reduction. This incremental selection repeats until the classification error begins to increase, i.e. $e_{k+1} > e_k$. Note that we allow the incremental search to continue when e_{k+1} equals e_k, because we want to search a space as large as possible. Once the termination condition is satisfied, the selected number of features, m, is chosen as the dimension for which the lowest error is first reached. For example, suppose the sequence of classification errors of the first 6 features is [10, 8, 4, 4, 4, 7], the forward selection will terminate at 5 features, but only return the first 3 features as the result; in this way we obtain a more compact set of features that minimizes the error.

3.3 Characteristic Feature Space

Given two feature sets S_n^1 and S_n^2 both containing n features, and a classifier Γ, we say the feature space of S_n^1 is more *characteristic* if the classification error (using classifier Γ) on S_n^1 is smaller than on S_n^2. This definition of characteristic space can be extended recursively to the subsets (subspaces) of S_n^1 and S_n^2. Suppose we have a feature selection method F to generate a series of feature subsets in S_n^1: $S_1^1 \subset S_2^1 \subset \ldots \subset S_k^1 \subset \ldots \subset S_{n-1}^1 \subset S_n^1$, and similarly a series of subsets in S_n^2: $S_1^2 \subset \ldots \subset S_k^2 \subset \ldots \subset S_n^2$. We say S_n^1 is *recursively more characteristic* (RM-characteristic) than S_n^2 on the range $\Omega = [k_{lower}, k_{upper}]$ ($1 \le k_{lower} < k_{upper} \le n$), if for every $k \in \Omega$, the classification error on S_k^1 is consistently smaller than on S_k^1.

To determine which one of the feature sets S_n^1 and S_n^2 is superior, it is often insufficient to compare the classification errors for a specific size of the feature sets. A better way is to observe which set is RM-characteristic for a reasonably large range Ω. In the extreme case, we use $\Omega = [1, n]$. Given two feature selection methods F^1 and F^2, if the feature sets generated by F^1 are RM-characteristic than those generated by F^2, we believe the method F^1 is better than F^2.

Let's consider the following example to compare mRMR and Max-Relevance based on the concept of RM-characteristic feature space. As a comprehensive study, we consider both the sequential and non-sequential feature sets as follows (more details will be given in experiments).

1) A direct comparison is to examine whether the mRMR sequential feature sets are RM-characteristic than Max-Relevance sequential feature sets. We use both methods to select n sequential feature sets $S_1 \subset \ldots \subset S_k \subset \ldots \subset S_n$ and compute the respective classification errors. If for most $k \in [1, n]$ we obtain smaller errors on mRMR feature sets, we can conclude that mRMR is better than Max-Relevance for the sequential (or incremental) feature selection.
2) We also use other feature selection methods (e.g. wrappers) in the second stage of our feature-selection algorithm to probe whether mRMR is better than Max-Relevance for non-sequential feature sets. For example, for the mRMR and Max-Relevance candidate feature sets with n^* features, we use the backward-selection-wrapper to produce two series of feature sets with $k = n^*-1, n^*-2, \ldots, m$ features by removing some non-sequential features that are potentially redundant. Then the respective classification errors of these feature sets are computed. If for most k we find the mRMR non-sequential feature subset leads to lower error, we conclude the mRMR candidate feature set is (approximately) RM-characteristic than the Max-Relevance candidate feature set.
3) Both the forward and backward selections of wrapper are used. Different classifiers (as discussed later in Section 4.2) are also considered in wrappers. We use both mRMR and Max-Relevance methods to select the same number of candidate features, and compare the classification errors of the feature subsets thereafter selected by wrappers. If all the observations agree that the mRMR candidate feature set is RM-characteristic, we have high confidence that mRMR is a superior feature selection method.

4) Given two feature sets, if S_n^1 is RM-characteristic than S_n^2, then it is faithful to compare the lowest errors obtained for the subsets of S_n^1 and S_n^2.

Clearly, for feature spaces containing the same number of features, wrappers can be applied more effectively on the space that is RM-characteristic. This also indicates wrappers can be applied at a lower cost, by improving the characterizing strength of features and reducing the number of pre-selected features.

In real situations, it might not be possible to obtain $e_k^1 < e_k^2$ for every k in Ω. Hence, we can define a confidence score $0 \le \rho \le 1$ to indicate the percentage of different k values for which the $e_k^1 < e_k^2$ condition is satisfied. For example, when ρ=0.90 (90% k-values correspond to the $e_k^1 < e_k^2$ condition), it is safe to claim that S_n^1 is *approximately* RM-characteristic than S_n^2 on Ω. As can be seen in the experiments, usually this approximation is sufficient to compare two series of feature subsets.

4. Implementation Issues

Before presenting the experimental results in Section 5, we discuss two implementation issues regarding the experiments: 1) calculation of mutual information for both discrete and continuous data, and 2) multiple types of classifiers used in our experiments.

4.1 Mutual Information Estimation

We consider mutual information based feature selection for both discrete and continuous data. For discrete (categorical) feature variables, the integral operation in Eq. (1) reduces to summation. In this case, computing mutual information is straightforward, because both joint and marginal probability tables can be estimated by tallying the samples of categorical variables in the data.

However, when at least one of variables x and y is continuous, their mutual information $I(x;y)$ is hard to compute, because it is often difficult to compute the integral in the continuous space based on a limited number of samples. One solution is to incorporate data discretization as a preprocessing step. For some applications where it is unclear how to properly discretize the continuous data, an alternative solution is to use density estimation method (e.g. Parzen windows) to approximate $I(x;y)$, as suggested by earlier work in medical image registration [7] and feature selection [17].

Given N samples of a variable x, the approximate density function $\hat{p}(x)$ has the following form,

$$\hat{p}(x) = \tfrac{1}{N} \sum_{i=1}^{N} \delta(x - x^{(i)},\, h)\,, \tag{17}$$

where δ(.) is the Parzen window function as explained below, $x^{(i)}$ is the ith sample, h is the window width. Parzen proved that with the properly chosen δ(.) and h, the estimation $\hat{p}(x)$ can converge to the true density $p(x)$ when N goes to infinity [21]. Usually, δ(.) is chosen as the Gaussian window:

$$\delta(z,h) = \exp(-\frac{z^T \Sigma^{-1} z}{2h^2}) \Big/ \{(2\pi)^{d/2} h^d \mid \Sigma \mid^{1/2}\}\,, \tag{18}$$

where $z = x - x^{(i)}$, d is the dimension of the sample x, and Σ is the covariance of z. When $d = 1$, Eq. (17) returns the estimated marginal density; when $d = 2$, we can use Eq. (17) to estimate the density of bivariate variable (x,y), $p(x,y)$, which is actually the joint density of x and y. For the sake of robust estimation, for $d \ge 2$, Σ is often approximated by its diagonal components.

4.2 Multiple Classifiers

Our mRMR feature selection method does not convolve with specific classifiers. Therefore, we expect the features selected by this scheme have good performance on various types of classifiers. To test this, we consider three widely used classifiers, i.e. Naïve Bayes (NB), Support Vector Machine (SVM), and Linear Discrimant Analysis (LDA).

NB [20] is one of the oldest classifiers. It is based on the Bayes rule and assumes that feature variables are independent of each other given the target class. Given a sample $s = \{x_1, x_2, \ldots, x_m\}$ for m features, the posterior probability that s belongs to class c_k is

$$p(c_k \mid s) \propto \prod_{i=1}^{m} p(x_i \mid c_k), \tag{19}$$

where $p(x_i|c_k)$ is the conditional probability table (or densities) learned from examples in the training process. The Parzen-window density-approximation in Eqs. (17) and (18) can be used to estimate $p(x_i|c_k)$ for continuous features. Despite the conditional independence assumption, NB has been shown to have good classification performance for many real data sets, on par with many more sophisticated classifiers [20].

SVM [29][2] is a more modern classifier that uses kernels to construct linear classification boundary in higher dimensional spaces. We use the LIBSVM package [9], which supports both 2-class and multi-class classification.

As one of the earliest classifiers, LDA [30] learns a linear classification boundary in the input feature space. It can be used for both 2-class and multi-class problems.

5. Experiments

We tested our feature selection approach on two discrete and two continuous data sets. For these data sets, we used multiple ways to calculate the mutual information and tested the performance of the selected features based on three classifiers introduced above. In this way, we provided a comprehensive study on the performance of our feature selection approach under different conditions.

This section is organized as follows. After a brief introduction of data sets in section 5.1, we compare mRMR against Max-Dependency in terms of both feature selection complexity and feature classification accuracy in section 5.2. These results demonstrate the practical advantages of our mRMR scheme and provide a direct verification of the theoretical analysis in Section 2. Then, in sections 5.3 and 5.4 we show a detailed comparison of mRMR and Max-Relevance, the latter of which has been widely used in practice. We do not show the comparison of mRMR with Min-Redundancy since Min-Redundancy alone usually leads to poor classification (and is seldom used to select features in real applications). Due to the space limitation, in the following we always demonstrate our comprehensive study with the most representative results. For simplicity, we use MaxDep to denote Max-Dependency and MaxRel to denote Max-Relevance, throughout the figures, tables, and texts in this section.

5.1 Data Sets

The four data sets we used are shown in Table 1. They have been extensively used in earlier studies [1][13][26][27][5]. The first two data sets, HDR-MultiFeature (HDR) and Arrhythmia (ARR), are also available on the UCI machine learning archive [28]. The latter two, NCI and Lymphoma (LYM), are available on the respective authors' web sites. All the raw data are continuous. Each feature variable in the raw data was preprocessed to have zero mean-value and unit variance (i.e. transformed to their z-scores). To test our approaches on both discrete and continuous data, we discretized the first two data sets, HDR and ARR. The other two data sets, NCI and LYM, were directly used for continuous feature selection.

The data set HDR [6][14][13][28] contains 649 features for 2000 handwritten digits. The target class has 10 states, each of which has 200 samples. To discretize the data set, each feature variable was binarized at the mean value, i.e., it takes 1 if it is larger than the mean value and –1 otherwise. We selected and evaluated features using 10-fold Cross-Validation (CV).

The data set ARR [28] contains 420 samples of 278 features. The target class has 2 states with 237 and 183 samples, respectively. Each feature variable was discretized into 3 states at the positions $\mu \pm \sigma$ (μ is the mean value, and σ the standard deviation): it takes -1 if it is less than $\mu-\sigma$, 1 if larger than $\mu+\sigma$, and 0 if otherwise. We used 10-fold CV for feature selection and testing.

The data set NCI [26][27] contains 60 samples of 9703 genes; each gene is regarded as a feature. The target class has 9 states corresponding to different types of cancer; each type has 2 ~ 9 samples. Since the

sample number is small, we used the Leave-One-Out (LOO) CV method in testing.

The data set LYM [1] has 96 samples of 4026 gene features. The target class corresponds to 9 subtypes of the lymphoma. Each subtype has 2 ~ 46 samples. The sample numbers for these subtypes are highly skewed, which makes it a hard classification problem.

Note that the feature numbers of these data sets are large (e.g. NCI has nearly 10000 features). These data sets represent some real applications where expensive feature selection methods (e.g. exhaustive search) cannot be used directly. They differ greatly in sample size, feature number, data type (discrete or continuous), data distribution, and target class type (multiclass or 2-class). In addition, we studied different mutual information calculation schemes for both discrete and continuous data, provided results using different classifiers and different wrapper selection schemes. We believe these data and methods provide a comprehensive testing suit for feature selection methods under different conditions.

5.2 Comparison of mRMR and Max-Dependency

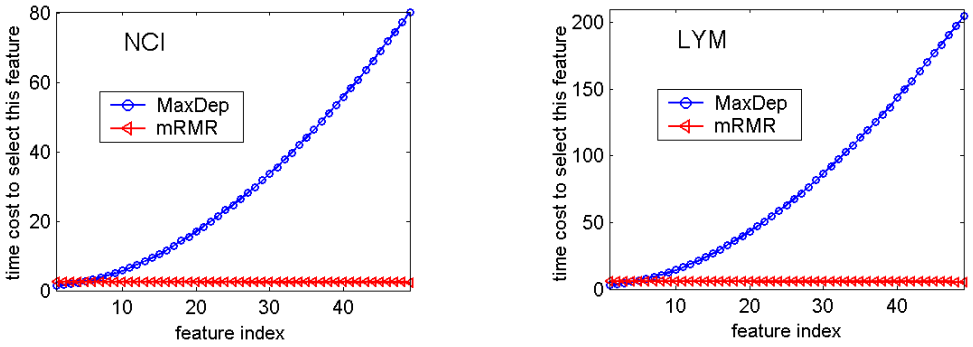
The mRMR scheme is a first-order approximation of the Max-Dependency (or MaxDep) selection method. We compared their performances in terms of both feature selection complexity and feature classification accuracy. These comparisons indicate their applicability for real data.

Table 1. Data sets used in our experiments

Data set	HDR MultiFeat		Arrhythmia		NCI		Lymphoma	
Acronym	HDR		ARR		NCI		LYM	
Source	UCI [28], Duin et al [6][14][13]		UCI [28]		Ross et al [26] Scherf et al [27]		Alizadeh et al [1]	
Raw data type	Continuous							
Experimental data type	Discrete				Continuous			
Processing method	Binarize at μ		Discretize at $\mu\pm\sigma$ to be 3-state		z-score (mean value 0, standard deviation 1)			
# Variable	649		278		9703		4026	
# Sample	2000		420		60		96	
# Class	10		2		9		9	
Class	Name	# Sample	Name	# Sample	Name	# Sample	Name	# Sample
C1	0	200	Normal	237	NSCLC	9	DLBCL	46
C2	1	200	Abnormal	183	Renal	9	CLL	11
C3	2	200			Breast	8	ABB	10
C4	3	200			Melanoma	8	FL	9
C5	4	200			Colon	7	RAT	6
C6	5	200			Leukemia	6	TCL	6
C7	6	200			Ovarian	6	RBB	4
C8	7	200			CNS	5	GCB	2
C9	8	200			Prostate	2	LNT	2
C10	9	200						
Testing method	10-fold CV				LOOCV			

5.2.1 Feature Selection Complexity

In practice, for categorical feature variables, we can introduce an intermediate "joint-feature" variable for MaxDep selection, so that the complexity would not increase much in selecting additional features (the comparison results against mRMR are omitted due to space limitation). Unfortunately, for continuous feature variables, it is hard to adopt a similar approach. For example, we compared the average computational time cost to select the top 50 mRMR and MaxDep features for both continuous data sets NCI and LYM, based on parallel experiments on a cluster of eight 3.06G Xeon CPUs running Redhat Linux 9, with the Matlab implementation.



(a) Time cost for selecting each NCI feature *(b) Time cost for selecting each LYM feature*

Fig. 1 Time cost (seconds) for selecting individual features based on mutual information estimation for continuous data sets.

The results in Fig. 1 demonstrate that the time cost for MaxDep to select a single feature is a polynomial function of the number of features, whereas for mRMR it is almost constant. For example, for NCI data, MaxDep takes about 20 and 60 seconds to select the 20th and 40th features, respectively. In contrast, mRMR always takes about 2 seconds to select any features. For the LYM data, MaxDep needs more than 200 seconds to find the 50th feature, while mRMR uses only 5 seconds. We can conclude that mRMR is computationally much more efficient than MaxDep.

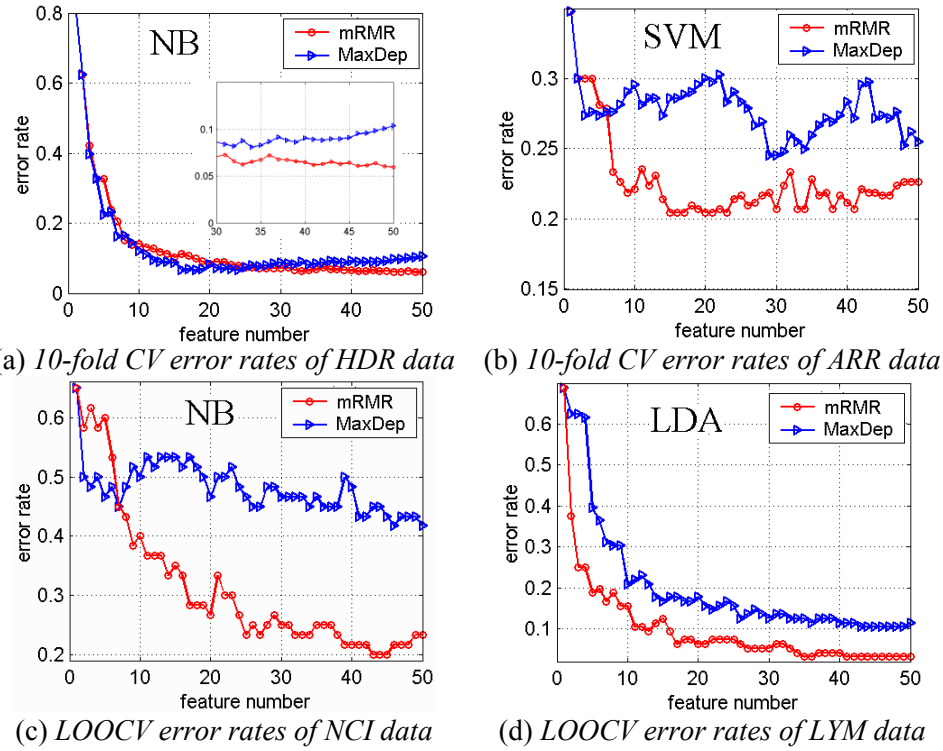
5.2.2 Feature Classification Accuracy

The selected features for the four data sets were tested using all the three classifiers introduced in section 4.2. However, for both clarity and briefness, we only plot several representatives of the cross-validation classification error-rate curves in Fig.2. Similar results were obtained in other cases.

Fig. 2 (a) shows that for the HDR data, the overall performance of MaxDep and mRMR is similar. MaxDep gets slightly lower errors when the feature number is relatively small, within the range between 1 and 20. When the feature number is larger than 30, the MaxDep features lead to a significantly greater error rate than mRMR features, as indicated in the blow-up windows. For example, 50 mRMR features lead to 6% error, in contrast to the 11% error of 50 MaxDep features. Noticeably, the two error-rate curves have distinct tendency. For mRMR, the error rate constantly decreases and then converges at some point. On the contrary, the error rate for MaxDep declines for small feature-numbers and then starts to increase for greater feature-numbers, indicating that more features lead to worse classification.

Fig. 2 (b) ~ (d) show the respective comparison results for ARR, NCI, and LYM data sets. The different tendency of the mRMR and MaxDep error-rate curves can be seen more prominently for these three data sets. For example, in (b) we see that only with the first 3~5 features, MaxDep has a slightly lower error rate than mRMR; but the respective error rates are far away from optimum. For all the rest feature numbers, mRMR features lead to consistently lower errors than MaxDep. For NCI data in (c), MaxDep is better than mRMR only when less than 7 features are used, but its overall classification accuracies are very poor. For a larger feature number, mRMR features lead to only half of the error rate of MaxDep features, indicating a greater discriminating strength. For LYM data in (d), MaxDep features are never better than mRMR features.

Why mRMR tends to outperform MaxDep when the feature number is relatively large? This is because in high dimensional space the estimation of mutual information becomes much less reliable than in 2-dimensional space, especially when the number of data samples is comparatively close to the number of joint states of features. This phenomenon is seen more clearly for continuous feature variables, i.e. the NCI and LYM data sets.



(a) *10-fold CV error rates of HDR data* (b) *10-fold CV error rates of ARR data*

(c) *LOOCV error rates of NCI data* (d) *LOOCV error rates of LYM data*

Fig. 2 Comparison of feature classification accuracies of mRMR and MaxDep. Classifiers NB, SVM, LDA were used.

This also explains why for HDR data, the difference between mRMR and MaxDep is not as prominent as those of the three other data sets. Because the HDR data set has a much larger number of data samples than ARR, NCI, and LYM data sets, the accuracy of mutual information estimation for HDR data does not degrade as quickly as those for the other three data sets.

Since the complexity of MaxDep in selecting features is higher and the classification accuracy using MaxDep features is lower, it is much more appealing to make use of mRMR instead of MaxDep in practical feature selection applications. In the following subsections, we will focus on comparing mRMR against the most widely used MaxRel selection method.

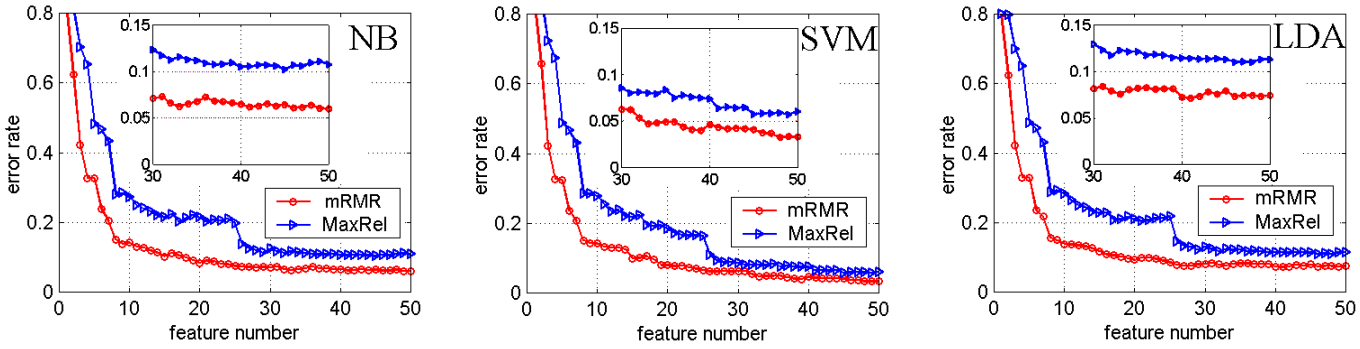
5.3 Comparison of Candidate Features Selected by mRMR and MaxRel

MaxRel and mRMR have similar computational complexity. The mRMR method is a little bit more expensive, but the difference is minor. Thus, we focus on comparing the feature classification accuracies.

5.3.1 Discrete Data

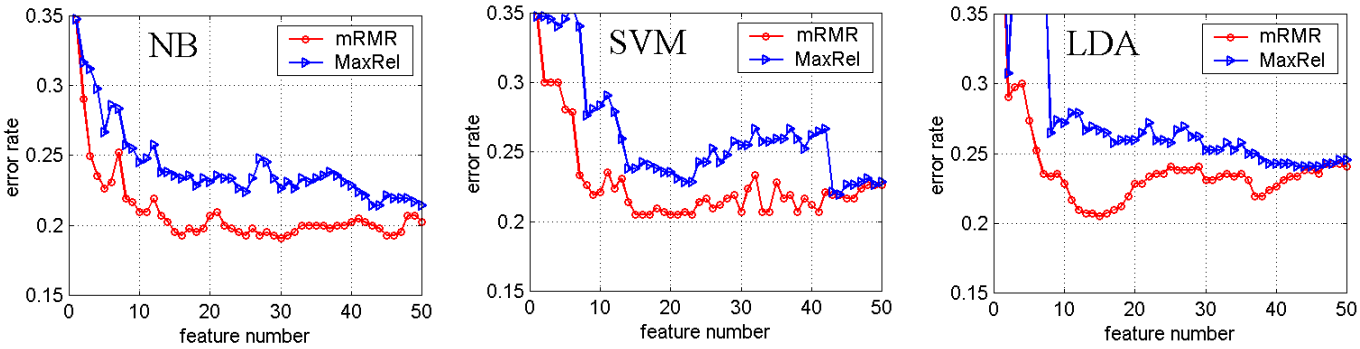
Figs. 3 and 4 show results of the incremental feature selection and classification for discrete data sets. The feature number ranges from 1 to 50.

For HDR data set, Fig. 3 (a)~(c) show the classification error rates with classifiers NB, SVM, and LDA respectively. Clearly, features selected by mRMR consistently attain significantly lower error rates than those selected by MaxRel. In other words, feature sets selected by mRMR are RM-characteristic than those selected by MaxRel. In this case, it is faithful to compare the lowest classification errors obtained for both methods. As illustrated in the zoom-in windows, with NB, the lowest error rate of mRMR is about 6%, while that of MaxRel is about 10%; with SVM, the lowest error rate of mRMR is about 3.5%, the lowest error rate of MaxRel is about 5.5%; with LDA, the lowest error rate of mRMR features is around 7%, whereas that of MaxRel is around 11%.



(a) NB classifier error rate *(b) SVM classifier error rate* *(c) LDA classifier error rate*

Fig. 3 10-fold CV error rates of HDR data using mRMR and MaxRel features.



(a) NB classifier error rate *(b) SVM classifier error rate* *(c) LDA classifier error rate*

Fig. 4 10-fold CV error rates of ARR data using mRMR and MaxRel features.

Fig. 4 (a)~(c) show the classification error rates for the ARR data. Similar to those of the HDR data, features selected by mRMR significantly and consistently outperform those selected by MaxRel. When nearly 50 features are used, the performance of mRMR and MaxRel become close. Overall, the performance of mRMR is much better than that of MaxRel, since 15 mRMR features lead to better classification accuracy than 50 MaxRel features.

Results in this section show that for discrete data sets, the candidate features selected by mRMR are significantly better than those selected by MaxRel. These effects are independent of the concrete classifiers we used.

5.3.2 Continuous Data

Tables 2 ~ 3 show the results of the incremental feature selection and classification for continuous data sets. The feature number ranges from 1 to 50 (to save space, we only list results of 1,5,10,15,…,50 features).

Table 2 shows that for NCI data, features selected by mRMR lead to lower error rates than those selected by MaxRel. The differences are consistent and significant. For example, with NB and more than 40 features (for simplicity, this combination is called "NB+40features"), we obtained an error rate around 20% for mRMR and around 33% for MaxRel. With SVM+40features, we obtained the error rate 23~26% for mRMR, and 35~38% for MaxRel. With LDA+40features, the results are similar.

Table 3 shows that for LYM data, mRMR features are also superior to MaxRel features (e.g. 3% versus 15% for LDA+50features).

Results in this section show that for continuous data, mRMR also outperforms MaxRel in selecting RM-characteristic sequential feature sets. They also indicate that the Parzen-window-based density-estimation for mutual information computation can be effectively used for feature selection.

Table 2. LOOCV error rate (%) of NCI data using mRMR and MaxRel features.

Classifier	*m* / Method	1	5	10	15	20	25	30	35	40	45	50
NB	MaxRel	65.00	51.67	51.67	45.00	46.67	43.33	41.67	38.33	36.67	33.33	36.67
	mRMR	65.00	60.00	40.00	35.00	26.67	23.33	25.00	25.00	21.67	20.00	23.33
SVM	MaxRel	98.33	46.67	55.00	50.00	45.00	55.00	41.67	35.00	38.33	35.00	36.67
	mRMR	98.33	70.00	58.33	48.33	40.00	31.67	31.67	31.67	26.67	23.33	23.33
LDA	MaxRel	73.33	60.00	60.00	50.00	46.67	46.67	41.67	36.67	38.33	41.67	40.00
	mRMR	73.33	66.67	50.00	53.33	45.00	33.33	35.00	35.00	33.33	30.00	30.00

Table 3. LOOCV error rate (%) of LYM data using mRMR and MaxRel features.

Classifier	*m* / Method	1	5	10	15	20	25	30	35	40	45	50
NB	MaxRel	72.92	25.00	15.63	13.54	13.54	12.50	13.54	12.50	11.46	11.46	10.42
	mRMR	72.92	17.71	16.67	10.42	11.46	9.38	10.42	9.38	9.38	7.29	8.33
SVM	MaxRel	42.71	27.08	21.88	21.88	18.75	16.67	14.58	14.58	15.63	11.46	12.50
	mRMR	42.71	11.46	10.42	7.29	5.21	7.29	7.29	5.21	5.21	5.21	4.17
LDA	MaxRel	68.75	32.29	22.92	23.96	23.96	21.88	22.92	17.71	16.67	16.67	15.63
	mRMR	68.75	18.75	15.63	12.50	6.25	7.29	5.21	3.13	4.17	3.13	3.13

5.4 Comparison of Compact Feature Subsets Selected by mRMR and MaxRel

The results of incrementally selected candidate features have indicated that with the same number of *sequential* features, mRMR feature set has more characteristic strength than MaxRel feature set. Here we investigate that given the same number of candidate features, whether the mRMR feature space is RM-characteristic and contains a more characterizing *non-sequential* feature subspace than the MaxRel feature space. This can be examined using the wrapper methods in Section 3.

Since the first 50 features lead to reasonably stable and small error for every data set and classification method we tested (see Figs. 3~4 and Tables 2~3), we used the first 50 features selected by MaxRel and mRMR as the candidate features.

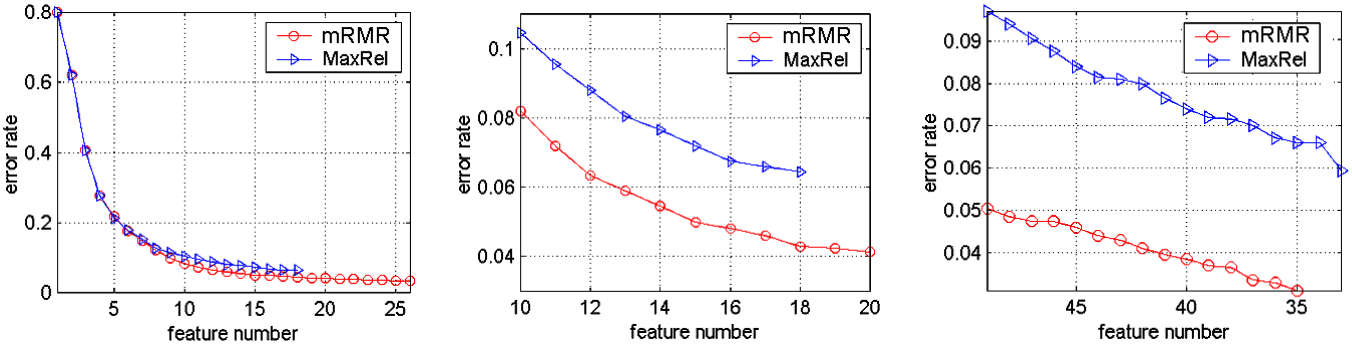
Both forward and backward selection wrappers were used to search for the optimal subset of features. If the candidate feature space of mRMR is RM-characteristic than that of MaxRel, wrappers should be able to find combinations of mRMR features that correspond to better classification accuracy.

As an example, Fig. 5 shows the classification error rates of optimal feature subsets selected by wrappers, for HDR data set and NB classifier. Fig.5 (a) and (b) (the zoom-in view of (a)) clearly show that forward-selection-wrapper can consistently find a significant better subset of features from the mRMR candidate feature set than from the MaxRel candidate feature set. This indicates mRMR candidate feature set is RM-characteristic for the forward-selection of wrapper. For MaxRel, wrapper obtains the lowest error 6.45% by selecting 18 features; more features will increase the error (thus the wrapper selection is terminated). In contrast, by selecting 18 mRMR features, wrapper has ~ 4% classification error; it achieves even lower classification error with more mRMR features, e.g. 3.2% error for 26 features.

Fig.5 (c) shows that the backward-selection-wrapper also finds superior subsets from the candidate features generated by mRMR. Such feature subset always leads to significantly lower error rate than the

subset selected from MaxRel candidate features. This indicates the space of candidate features generated by mRMR does embed a subspace in which the data samples can be more easily classified.

Table 4 summarizes the results obtained for all four data sets and three classifiers. Obviously, similar to the HDR data, for almost all combinations of data sets, wrapper selection methods, and classifiers, lower error rates are attained from the mRMR candidate features, indicating that wrappers find more characterizing feature subspaces from the mRMR features than from MaxRel features. We can conclude that mRMR candidate feature sets do cover a wider spectrum of the more characteristic features. (There are two exceptions in Table 4, for which the obtained feature subsets are comparable: 1) "NCI+LDA+Forward", where 5 mRMR features lead to 20 errors (33.33%) and 7 MaxRel features lead to 19 errors (31.67%), and 2) "LYM+SVM+Backward", the same error (3.13%) is obtained).



(a) Forward selection (b) Forward selection (zoom-in) (c) Backward selection

Fig. 5 The wrapper selection/classification results (HDR + NB)

Table 4. Comparison of different wrapper selection results (lowest error rate (%))

Data set	Wrapper	NB		SVM		LDA	
		MaxRel	mRMR	MaxRel	mRMR	MaxRel	mRMR
HDR	Forward	6.45	3.20	5.50	3.45	6.80	4.05
	Backward	5.95	3.10	4.55	2.85	6.90	4.00
ARR	Forward	18.81	17.86	20.95	19.29	19.76	18.10
	Backward	23.10	17.86	20.48	19.52	20.48	18.33
NCI	Forward	26.67	13.33	25.00	21.67	31.67	33.33
	Backward	20.00	15.00	18.33	13.33	30.00	25.00
LYM	Forward	6.25	5.21	6.25	2.08	6.25	2.08
	Backward	5.21	3.13	3.13	3.13	6.25	3.13

6. Discussions

In our approach, we have stressed that a well-designed filter method, such as mRMR, can be used to enhance the wrapper feature selection, in achieving both high accuracy and fast speed. Our method uses an optimal first-order incremental selection to generate a candidate list of features that cover a wider spectrum of characteristic features. These candidate features have similar generalization strength on different classifiers (as seen in Figs. 3~4 and Tables 2~3). They facilitate effective computation of wrappers to find compact feature subsets with superior classification accuracy (as shown in Fig. 5 and Table 4). Our algorithm is especially useful for large-scale feature/variable selection problems where there are at least thousands of features/variables, such as medical morphometry [23][8], gene selection [32][31][5][12], etc.

Of note, the purpose of mRMR approach studied in this paper is to maximize the dependency. This typically involves the computation of multivariate joint probability, which is nonetheless difficult and inaccurate. Combining both Max-Relevance and Min-Redundancy criteria, the mRMR incremental selection scheme provides a better way to maximize the dependency. In this case, the difficult problem of multivariate joint probability estimation is reduced to estimation of multiple bivariate probabilities (densities), which is much easier. Our comparison in section 5.2 demonstrates that mRMR is a very good approximation scheme to Max-Dependency. In most situations, mRMR reduces the feature selection time dramatically for continuous features and improves the classification accuracy significantly. For data sets with a large number of samples, e.g. the HDR data set, the classification accuracy of mRMR is close to or better than that of Max-Dependency. We notice that the mRMR approach could also be applied to other domains where the similar heuristic algorithms are applicable to maximize the dependency of variables, such as searching (learning) the locally optimal structures of Bayesian networks [24].

Our scheme of mRMR does not intend to select features that are independent of each other. Instead, at each step, it tries to select a feature that minimizes the redundancy and maximizes the relevance. For real data, the features selected in this way will have more or less correlation with each other. However, our analysis and experiments show that the joint effect of these features can lead to very good classification accuracy. A set of features that are completely independent of each other usually would be less optimal.

All the feature selection methods used in this paper, including incremental search, forward or backward selection, etc., are heuristic search methods. None of them can guarantee the global maximization of a criterion function. The fundamental problem is the difficulty in searching the whole space, as pointed out at the beginning of this paper. Additionally, questing the global optimum strictly might lead to data overfitting. On the contrary, mRMR seems to be a practical way to achieve superior classification accuracy in relatively low computational complexity.

Our experimental results show that although in general more mRMR features will lead to a smaller classification error, the decrement of error might not be significant for each additional feature, or occasionally there could be fluctuation of classification errors. For example, in Fig. 3, the fifth mRMR feature seemingly has not led to a major reduction of the classification error produced with the first four features. Many factors count for these fluctuations. One cause is that additional features might be noisy. Another possible cause is that the mRMR scheme in Eq. (6) takes difference of the relevance term and the redundancy term. It is possible that one redundant feature also has relatively large relevance, so it could be selected as one of the top features. A greater penalty on the redundancy term would lessen this problem. A third possible cause is that the cross-validation method used might also introduce some fluctuations of the error curve. While a more detailed discussion on this fluctuation problem and other potential causes is beyond the scope of this paper, a way to solve this problem is to use other feature selectors to directly minimize the classification error and remove those potentially unneeded features, as what we do in the second-stage of our algorithm. For example, by using wrappers in the second stage, the error curves thus obtained in Fig. 5 are much smoother than those obtained using the first-stage only in Fig. 3.

Our results show that for continuous data, density estimation works well for both mutual information calculation and naïve Bayes classifier. In earlier work, Kwak et al [17] used the density estimation approach to calculate the mutual information between an individual feature x_i and the target class c. We used a different approach based on direct Parzen-window approximation similar to [7]. Our results indicate that the estimated density and mutual information among continuous feature variables can be utilized to reduce the redundancy and improve the classification accuracy.

Finally, we notice that Eq. (6) is not the only possible mRMR scheme. Instead of combining relevance and redundancy terms using difference, we can consider quotient [5] or other more sophisticated schemes. The quotient-combination imposes a greater penalty on the redundancy. Empirically, it often leads to better classification accuracy than the difference-combination for candidate features. However, the joint effect of these features is less robust when some of them are eliminated; as a result, in the second-stage of feature selection using wrappers, the set of features induced from the quotient-combination would have a bigger size than that from the difference-combination. The mRMR

paradigm can be better viewed as a general framework to effectively select features and allow all possibilities for more sophisticated or more powerful implementation schemes.

7. Conclusions

We present a theoretical analysis of the minimal-redundancy-maximal-relevance (mRMR) condition and show that it is equivalent to the maximal dependency condition for first-order feature selection. Our mRMR incremental selection scheme avoids the difficult multivariate density estimation in maximizing dependency. We also show that mRMR can be effectively combined with other feature selectors such as wrappers to find a very compact subset from candidate features at lower expense. Our comprehensive experiments on both discrete and continuous data sets and multiple types of classifiers demonstrate that the classification accuracy can be significantly improved based on mRMR feature selection.

Acknowledgments

We thank Chris Ding for helpful discussion on mRMR and the density estimation methods, Hao Chen for discussion on heuristic search schemes, Yu Wang for comments on a preliminary version of this paper, and the anonymous reviewers for their helpful comments.

References

[1] Alizadeh, A.A., et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," Nature, vol.403, pp.503-511, 2000.

[2] Burges, C.J.C., "A tutorial on support vector machines for pattern recognition," Knowledge Discovery and Data Mining, vol.2, pp.1-43, 1998.

[3] Cover, T., and Thomas, J., *Elements of Information Theory*, New York: Wiley, 1991.

[4] Cover, T.M., "The best two independent measurements are not the two best," IEEE Trans. Systems, Man, and Cybernetics, vol. 4, pp. 116-117, 1974.

[5] Ding, C., and Peng, H.C., "Minimum redundancy feature selection from microarray gene expression data," Proc. 2nd IEEE Computational Systems Bioinformatics Conf., pp.523-528, Stanford, CA, Aug, 2003.

[6] Duin, R.P.W., and Tax, D.M.J., Experiments with Classifier Combining Rules, in: J. Kittler, F. Roli (eds.), *Multiple Classifier Systems* (Proc. First Int Workshop, MCS 2000, Cagliari, Italy, June 2000), Lecture Notes in Computer Science, vol. 1857, Springer, Berlin, pp. 16-29, 2000.

[7] Hadley, S.W., Pelizzari, C., and Chen, G.T.Y., "Registration of localization images by maximization of mutual information," AAPM 1996.

[8] Herskovits, E., Peng, H.C., and Davatzikos, C., "A Bayesian morphometry algorithm," IEEE Trans. on Medical Imaging, vol. 24, no. 6, pp. 723-737, 2004

[9] Hsu, C.W., and Lin, C.J., "A comparison of methods for multi-class support vector machines," IEEE Transactions on Neural Networks, vol.13, pp.415-425, 2002.

[10] Hyvärinen, A., Karhunen, J., & Oja, E., *Independent Component Analysis*, John Wiley & Sons, 2001.

[11] Iannarilli, F.J., and Rubin, P.A., "Feature selection for multiclass discrimination via mixed-integer linear programming," IEEE Trans. Pattern Analysis and Machine Intelligence, 25(6), pp.779-783, 2003.

[12] Jaeger, J., Sengupta, R., and Ruzzo, W.L., "Improved gene selection for classification of microarrays," PSB'2003, pp.53-64, 2003.

[13] Jain, A.K., and Zongker, D., "Feature selection: evaluation, application, and small sample performance," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 2, pp. 153-158, Feb, 1997.

[14] Jain, A.K., Duin, R.P.W., and Mao, J., "Statistical pattern recognition: a review," IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(1): 4-37, 2000.

[15] Kohavi, R., and John, G., "Wrapper for feature subset selection," Artificial Intelligence, vol. 97, no.1-2, pp.273-324, 1997.

[16] Krantz, S. G., "Jensen's Inequality," §9.1.3 in *Handbook of Complex Analysis*. Boston, MA: Birkhäuser, pp. 118, 1999.

[17] Kwak, N., and Choi, C.H., "Input feature selection by mutual information based on Parzen window," IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(12), pp. 1667-1671, 2002.

[18] Langley, P., "Selection of relevant features in machine learning," AAAI Fall Symposium on Relevance, 1994.

[19] Li, W., and Yang Y., "How many genes are needed for a discriminant microarray data analysis?" Critical Assessment of Techniques for Microarray Data Mining Workshop, Dec 2000. pp. 137-150.

[20] Mitchell, T., Machine Learning. McGraw-Hill. 1997.
[21] Parzen, E., "On estimation of a probability density function and mode," *Annals of Math. Statistics*, vol. 33, pp. 1065-1076, 1962.
[22] Peng H.C., Gan, Q., and Wei, Y., "Two optimization criterions for neural networks and their applications in unconstrained character recognition," Journal of Circuits and Systems (in Chinese), Vol.2, No.3, pp.1-6, 1997.
[23] Peng H.C., Herskovits E.H., and Davatzikos C., "Bayesian clustering methods for morphological analysis of MR images," Int. Symp. on Biomedical Imaging: From Nano to Macro, Washington, D.C., pp.485-488, 2002.
[24] Peng, H.C., and Ding, C., "Structural search and stability enhancement of Bayesian networks," Proc. 3rd IEEE Int. Conf. Data Mining, pp.621-624, Melbourne, Florida, USA, Nov, 2003.
[25] Pudil, P., Novovicova, J., and Kittler, J., "Floating search methods in feature selection," Pattern Recognition Letters, vol. 15, no. 11, pp. 1,119-1,125, 1994.
[26] Ross, D.T., Scherf, U., et al, "Systematic variation in gene expression patterns in human cancer cell lines," Nature Genetics, vol.24, no.3, pp.227-234, 2000.
[27] Scherf, U., Ross, D.T., et al, "A cDNA microarray gene expression database for the molecular pharmacology of cancer," Nature Genetics, vol.24, no.3, pp.236-244, 2000.
[28] UCI Machine Learning Repository: http://www.ics.uci.edu/~mlearn/MLSummary.html
[29] Vapnik, V., *The Nature of Statistical Learning Theory*, Springer: New York, 1995.
[30] Webb, A., *Statistical Pattern Recognition*, Arnold, 1999.
[31] Xing, E.P., Jordan, M.I., and Karp, R.M., "Feature selection for high-dimensional genomic microarray data," ICML2001, 2001.
[32] Xiong, M., Fang, Z., and Zhao, J., "Biomarker identification by feature wrappers," Genome Research, vol.11, pp.1878-1887, 2001.